# A. Appendix

## A.1. Data Collection

### A.1.1 Rig

To accommodate SANPO-Real's multi stereo camera requirements we designed a specialized data collection rig. This rig prioritizes hardware integration, reliable GPU cooling, and comfort for the wearer. Our setup involves volunteers wearing head and chest mounted ZED cameras (ZED-M and ZED-2i, respectively), with supporting hardware in a backpack. We also developed a mobile app for visualization and to control the data collection. Figure 9 shows the data collection system in action.

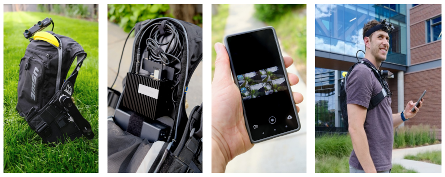

Figure 9. **SANPO-Real Data Collection Rig.**

## A.2. Dataset

### A.2.1 SANPO-Synthetic Reproducibility and rendering environment.

We created SANPO-Synthetic through our collaboration with a third party, Parallel Domain. If other researchers wish to reproduce these environments with other tools (NVidia Omniverse, Unreal, Unity, etc), we would welcome that. To aid reproducibility, here are detailed specifications for SANPO-Synthetic's virtual rendering environment.

All % are at session level.

1. Scene types : Urban environments only.
2. Camera Type : Zed 2i
   - (a) Image width: 2208
   - (b) Image height: 1242
   - (c) fx: 1914.203
   - (d) fy: 1914.203
   - (e) cx: 1074.4403
   - (f) cy: 655.79846
   - (g) camera matrix:

$$\begin{pmatrix} 1914.203 & 0 & 1074.4403 \\ 0 & 1914.203 & 655.79846 \\ 0 & 0 & 1 \end{pmatrix}$$

   - (h) stereo transform (between left/right cameras):

$$\begin{pmatrix} 1 & 0 & 0 & 119.96817 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

3. Camera Type : Zed Mini
   - (a) image_width: 2208
   - (b) image_height: 1242
   - (c) fx: 1376.4702
   - (d) fy: 1376.4702
   - (e) cx: 1112.7797
   - (f) cy: 599.8397
   - (g) camera matrix:

$$\begin{pmatrix} 1376.4702 & 0 & 1112.7797 \\ 0 & 1376.4702 & 599.8397 \\ 0 & 0 & 1 \end{pmatrix}$$

   - (h) stereo transform (between left/right cameras):

$$\begin{pmatrix} 1 & 0 & 0 & 62.944813 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

4. Camera Positions
   - (a) Zed 2i on chest.
   - (b) Zed mini just above head.
   - (c) Both with natural tilt variations.
   - (d) 50% of sessions from each position.
5. FPS (Frames per second)
   - (a) 60% at 5 FPS.
   - (b) 20% at 14.28 FPS.
   - (c) 20% at 33.33 FPS.
6. Ground truth annotations
   - (a) Panoptic segmentation mask.
   - (b) Metric depth map.
7. Lighting and Weather
   - (a) 70% well lit sunny
   - (b) 10% are at dawn/dusk with the sun low in the horizon
   - (c) 10% are dark/nighttime
   - (d) 5% have fog

    (e) 5% have rain

8. Obstacles
    (a) Garbage can
        i. 50% One per street block.
        ii. 30% two garbage cans. E.g: One normal and one recycle.
        iii. 20% no garbage can.
    (b) Trash bags
        i. 50% None
        ii. 40% 1-2
        iii. 10% >=5
    (c) Bike racks : One per street block.
    (d) Mailbox
        i. 60% one per street block.
        ii. 20% two adjacent mailboxes per street block.
        iii. 20% None.
    (e) Fire Hydrant
        i. 80% One per street block.
        ii. 20% None.
    (f) Construction cones : As provided by the rendering scene map.

9. Road Vehicle : Low, mid and high is the setting in the rendering engine.
    (a) 20% None
    (b) 30% low
    (c) 30% mid
    (d) 20% high

10. Pedestrians
    (a) 10>= per street block (50%)
    (b) 5>= per street block (30%)
    (c) <3 per street block (20%)
    (d) 20% very close to the ego person.

11. Trees on sidewalk
    (a) 60% high density.
    (b) 20% low density.
    (c) 20% no trees on the sidewalk.

12. Other naturally occurring things like curbs, dips, crosswalks, parking meters, traffic signs and lights, fences, plants, hedges etc.. will be included as provided by the rendering scene map.

13. Not Supported
    (a) Bike paths.
    (b) Riders on sidewalk.
    (c) Foliage and seasonal color changes of leaves.

### A.2.2 Dataset Comparison

While many outdoor video datasets exist for tasks like robot navigation, autonomous driving, and video segmentation (see Table 6), SANPO fills a crucial gap. To our knowledge, it is the only dataset providing both real and synthetic data with panoptic labels and depth maps specifically designed for human-centric egocentric navigation research.

Figure 10 provides a visual comparison between SANPO-Real and SANPO-Synthetic.

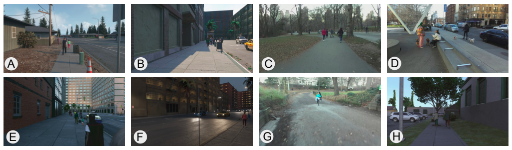

Figure 10. **SANPO Synthetic vs real.** A sample of SANPO-Real and SANPO-Synthetic data. *How quickly can you tell which of these images is synthetic?* Answer key in base64: ‘`c3ludGg6IEFCRUZILCByZWFsOiBDREc=`’

### A.2.3 Additional Statistics

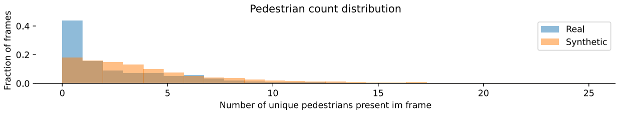

Figure 11. **Distribution of pedestrians** in SANPO-Real and SANPO-Synthetic. SANPO-Real frequently features images with no pedestrians, but pedestrians appear in almost all frames of SANPO-Synthetic, and in greater quantities.

#### Pedestrian density

#### Additional Attributes of SANPO-Synthetic’s Segmentation Annotations

- Instance Density: Over half of the frames have $\geq 60$ unique instances, with a sixth having $\geq 150$.
- Small Objects: 80% of object masks have less than $32^2$ pixels, significantly more than SANPO-Real (8.1%).

## A.3. Data Annotation

### A.3.1 Session Attributes

Each real session is annotated with the following high level attributes.

1. Human Traffic
    (a) Low
    (b) Moderate

| Dataset | Domain | Environment | # Frames | # Seg Masks | # Depth Maps |
|---|---|---|---|---|---|
| SCAND [20] | 🤖 | 🏠 + 🌲<br>👀 | ∼522<br>minutes | | ∼522<br>minutes |
| MuSoHu [34] | 🤖 | 🏠 + 🌲<br>👀 | ∼600<br>minutes | | ∼600<br>minutes |
| Playing for Benchmark (Synthetic) [41] | 🚗 | 🌲 | 250K | 250K<br>(Dense) | |
| Cityscapes-DVPS [39] | 🚗 | 🌲 + 👀 | 3K | 3K<br>(Dense) | 3K<br>(Dense) |
| KITTI-360 [26] | 🚗 | 🌲 | 320K | 2x78K<br>(Dense) | 2x78K<br>(Dense) |
| Panoptic-nuScenes [7] | 🚗 | 🌲 | 1.4M | 40K<br>(Dense) | |
| Waymo Open Dataset -Panoramic [30] | 🚗 | 🌲 | 390K | 100K | |
| A*3D [38] | 🚗 | 🌲 | 39K | 39K<br>(3D BBox) | |
| ApolloScape-SceneParsing [18] | 🚗 | 🌲 | 140K | 140K<br>(Dense) | 140K<br>(Dense) |
| DDAD [14] | 🚗 | 🌲 | 21K | | 21K<br>(Dense) |
| DOLPHINS (Synthetic) [29] | 🚗 | 🌲 | 42K | 42K<br>(3D BBox) | |
| Argoverse2 Sensor Data [45] | 🚗 | 🌲 + 👀 | 1000<br>(Videos) | 3D BBox | |
| CamVid [5] | 🚗 | 🌲 | 5<br>(Videos) | 700<br>(Dense) | |
| MS-COCO [27] | O.D.S | 🏠 + 🌲 | 328K | 328K | |
| Youtube-VOS [46] | V.O.S. | 🏠 + 🌲 | ∼20K | ∼4K<br>(Sparse) | |
| DAVIS-2017 [6] | V.O.S. | 🏠 + 🌲 | 10K | 10K<br>(Sparse) | |
| SideGuide [35] | 🚶 | 🌲 | 2x180K,<br>312K | 100K<br>(Sparse) | 180K<br>(Dense) |
| **SANPO-Real** (ours) | 🚶 | 🌲 + 👀 | 2x617K | 112K<br>(Dense) | 617K<br>(Dense) |
| **SANPO-Synthetic** (ours) | 🚶 | 🌲 | 113K | 113K<br>(Dense) | 113K<br>(Dense) |

🤖 : Robot Navigation, 🚗 : Self-Driving, 🚶 : Egocentric Navigation,
O.D.S: Object Detection & Segmentation, V.O.S: Video Object Segmentation
🏠 : Indoor, 🌲: Outdoor, 👀 Stereo

Table 6. **Dataset Comparison**. SANPO is a unique video dataset designed to address a gap in current offerings. Unlike existing datasets focused on self-driving vehicles or general video object segmentation (VOS), SANPO targets the specific challenges of egocentric human navigation. SANPO is a large-scale, challenging, and diverse dataset. It offers both real and synthetic data, with multi-view stereo data included in the real component.

(c) Heavy

2. Vehicular Traffic

   (a) Low

   (b) Moderate

   (c) Heavy

3. Animal Traffic

   (a) Low

   (b) Moderate

   (c) Heavy

4. Number of Obstacles

   (a) Low

   (b) Moderate

   (c) Heavy

5. Environment Type

   (a) Urban

   (b) Suburban

   (c) Rural

   (d) Park

   (e) Road Junction

   (f) Open Terrain

   (g) Open Space

   (h) Indoor

6. Weather Condition

   (a) Sunny

   (b) Cloudy

   (c) Rainy

   (d) Snowy

7. Visibility

   (a) High

   (b) Medium

   (c) Low

8. Motion Type

   (a) Walking

   (b) Jogging

   (c) Running

9. Elevation Change

   (a) Flat

   (b) Uphill

   (c) Downhill

   (d) Stairs

10. Ground Appearances

    (a) Light Gray

    (b) Dark Gray

    (c) Pavers

    (d) Color

    (e) Terrain

    (f) Gravel

    (g) Sand

11. Motion Blur

    (a) Low

    (b) Medium

    (c) High

12. Rare Events

### A.3.2 SANPO Taxonomy

SANPO taxonomy labels with *stuff* or *thing* distinction.

0. unlabeled → stuff
1. road → stuff
2. curb → stuff
3. sidewalk → stuff
4. guard rail/road barrier → stuff
5. crosswalk → thing
6. paved trail → stuff
7. building → stuff
8. wall/fence → stuff
9. hand rail → stuff
10. opening-door → thing
11. opening-gate → thing
12. pedestrian → thing
13. rider → thing
14. animal → thing
15. stairs → thing

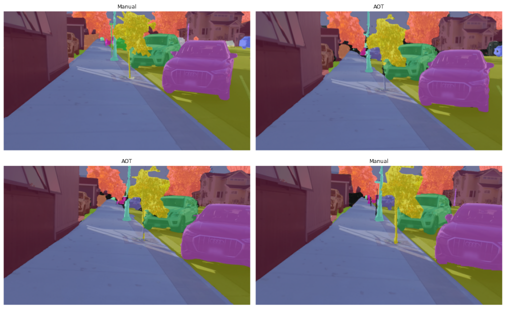

Figure 12. **Temporally Consistent Segmentation Annotation.** Our annotation process ensures temporal consistency across both human-annotated and machine-propagated masks. Compare the first and last columns of the figure to see this consistency. Most propagated masks are accurate, with occasional failures for thin objects like trees (yellow) and poles (cyan).

This process resulted in 18,787 human-annotated frames and 93,981 machine-propagated frames.

**Evaluating AOT-Based Propagation Accuracy** To evaluate the accuracy of AOT-based propagation for segmentation annotations, we performed the following analysis. We considered human-annotated frames (0-6-12-...), propagated segmentation masks to every other frame (6-18-...) using AOT, and compared these propagated masks to the corresponding human-annotated ground truth (GT) to calculate a propagation score. Since the motion gap between these frames is significant, this method provides a conservative estimate (lower bound) of the propagation error. In accordance with the video object segmentation (VOS) literature [10, 37, 49], we used region similarity $J$ and contour accuracy $F$ as evaluation metrics. The mean $J\&F$ score for SANPO-Real is 0.892, demonstrating a strong lower bound on the accuracy of machine-propagated masks.

**Detailed and Accurate Segmentation Annotation** Our dataset captures rich details, including high-quality semantic masks for even the smallest objects (see Fig. 13 for examples).

## A.4. Benchmarks

### A.4.1 Zero Shot Evaluation

**Cityscapes-19 -> SANPO Mapping**
To ensure a fair comparison, we map Cityscapes-19 labels to SANPO labels wherever possible. Below mapping from Cityscapes-19 to SANPO taxonomy:

1. road → road
2. sidewalk → sidewalk
3. building → building
4. wall → wall/fence

[3] We experimented with various combinations to refine this approach.

16. water body → stuff
17. other walkable surface → stuff
18. inaccessible surface → stuff
19. railway track → stuff
20. obstacle → thing
21. vehicle → thing
22. traffic sign → thing
23. traffic light → thing
24. pole → thing
25. bus stop → thing
26. bike rack → thing
27. sky → stuff
28. tree → thing
29. vegetation → stuff
30. terrain → stuff

### A.3.3 Segmentation Annotation Process

In this section we describe the segmentation annotation process for SANPO-Real. We divide each video into 30-second sub-videos (note: most videos are only 30 seconds long, resulting in a single sub-video), then we annotate every sixth frame (0-6-12-...), for a total of 90 frames per sub-video. To enhance efficiency and accuracy of human annotation, we employ two key techniques:

- Cascaded Annotation: To manage our extensive taxonomy, we divide all labels into five mutually exclusive subsets containing commonly co-occurring labels. Each sub-video is annotated in a temporally consistent manner across these subsets in a carefully determined optimal order[3]. When annotating a subset, previously annotated regions are frozen and displayed to the annotator, thus increasing their speed and improving boundary precision. The final subset includes all labels, ensuring that any regions missed in previous subsets are annotated.
- AOT based Propagation: We leverage AOT [49] to propagate masks from human-annotated frames to the intermediate unannotated frames. We track whether each frame is human-annotated or machine-propagated, and this information is included alongside the provided annotations. Figure 12 visually demonstrates this process, showing human-annotated frames and their machine-propagated counterparts.

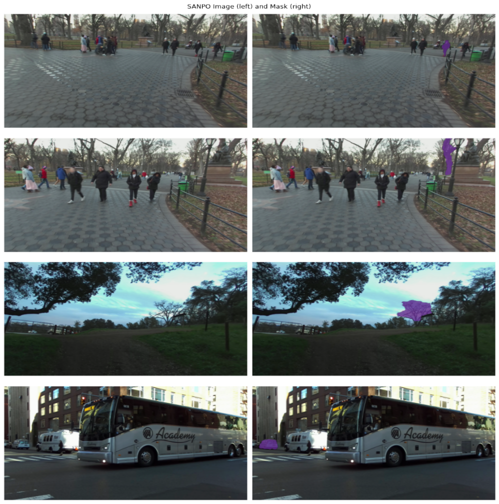

Figure 13. SANPO's detailed annotation include masks for even the smallest objects (highlighted in purple, right column).

5. fence → wall/fence
6. pole → pole
7. traffic light → traffic light
8. traffic sign → traffic sign
9. vegetation → vegetation
10. terrain → terrain
11. sky → sky
12. person → pedestrian
13. rider → rider
14. car → vehicle
15. truck → vehicle
16. bus → vehicle
17. train → vehicle
18. motorcycle → vehicle
19. bicycle → vehicle

For all SANPO labels without an appropriate mapping from Cityscapes-19, we treat the corresponding pixels as unlabeled and exclude them from the mIoU metric computation in the zero-shot semantic segmentation evaluation. The following SANPO labels were excluded:

1. curb
2. guard rail/road barrier
3. crosswalk
4. paved trail
5. hand rail
6. opening-door
7. opening-gate
8. animal
9. stairs
10. water body
11. other walkable surface
12. inaccessible surface
13. railway track
14. obstacle
15. bus stop
16. bike rack
17. tree

**Zero-shot Mask2Former Evaluation**

We also evaluated the Mask2Former Swin-L model [8] in the zero-shot setting. Despite its strong performance on Cityscapes (mIoU 0.833), it achieved lower scores on SANPO-Real (0.417) and SANPO-Synthetic (0.476). Fig. 14 offers a qualitative assessment on SANPO samples and Table 7 provides a class-wise mIoU breakdown.

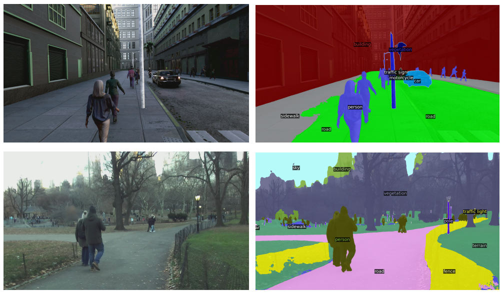

Figure 14. **Highlight on Domain Gap.** Egocentric navigation models must accurately differentiate between road (a not safe to walk surface) and sidewalk (a safe to walk surface). Mask2Former trained on Cityscapes dataset, similar to the Kmax-Deeplab models, struggles with this distinction on SANPO samples (top: synthetic, bottom: real). This, along with Table 1, underscores the limited transferability of such datasets to human-centric navigation tasks. This visualization is generated using the Mask2Former tool [8].

### A.4.2 SANPO Benchmark

To ensure fairness and reproducibility, we maintained the following training setup:

1. **Encoder Pretraining:** All encoders pretrained on ImageNet [42].
2. **Datasets Used:** Only SANPO train sets.
3. **Resizing:** Data resized to 1089x1921 (height x width), padding used to maintain aspect ratio.
4. **Hyperparameters:** Standard values as defined in [44].
5. **Training Budget:** 60,000 steps with a batch size of 32 (doubled for Synthetic-to-Real domain adaptation fine-tuning experiments (-> and + rows in Table 4)). Approximate epochs for reference:
    - Cityscapes Panoptic Segmentation: 645
    - SANPO-Real Panoptic Segmentation: 21
    - SANPO-Real (Human GT Only) Panoptic Segmentation: 129
    - SANPO-Synthetic Panoptic Segmentation and Depth Estimation: 21
    - SANPO-Real Depth Estimation: 4

| | mIoU | |
|---|---|---|
| Mapped SANPO Label | SANPO-Real | SANPO-Synthetic |
| road | 0.255 | 0.407 |
| sidewalk | 0.120 | 0.262 |
| building | 0.642 | 0.934 |
| wall/fence | 0.448 | 0.087 |
| pedestrian | 0.679 | 0.878 |
| rider | 0.271 | 0.247 |
| vehicle | 0.658 | 0.817 |
| traffic sign | 0.212 | 0.240 |
| traffic light | 0.127 | 0.344 |
| pole | 0.310 | 0.586 |
| sky | 0.658 | 0.919 |
| vegetation | 0.654 | 0.303 |
| terrain | 0.394 | 0.166 |
| **Average** | **0.417** | **0.476** |

Table 7. **Mask2Former Zero-Shot Evaluation:** Per label breakdown of mIoU on the Mask2Former (Cityscapes) zero-shot experiment.

## A.5. SANPO Dense Prediction Qualitative Examples

We show some example images in Fig. 15, as well as ground truth and predicted segmentation maps from kMax-Deeplab and ground truth and predicted depth maps from Binsformer.

## A.6. Application

### A.6.1 SANPO -> Accessibility Mapping

*"safe to walk"* (e.g. sidewalk) and *"not safe to walk"* (e.g. road, which is for vehicles) are ground surfaces.

1. unlabeled → not safe to walk
2. road → not safe to walk
3. curb → not safe to walk
4. sidewalk → safe to walk
5. guard rail/road barrier → obstacle
6. crosswalk → safe to walk
7. paved trail → safe to walk
8. building → obstacles
9. wall/fence → obstacles
10. hand rail → obstacles
11. opening-door → obstacles
12. opening-gate → obstacles

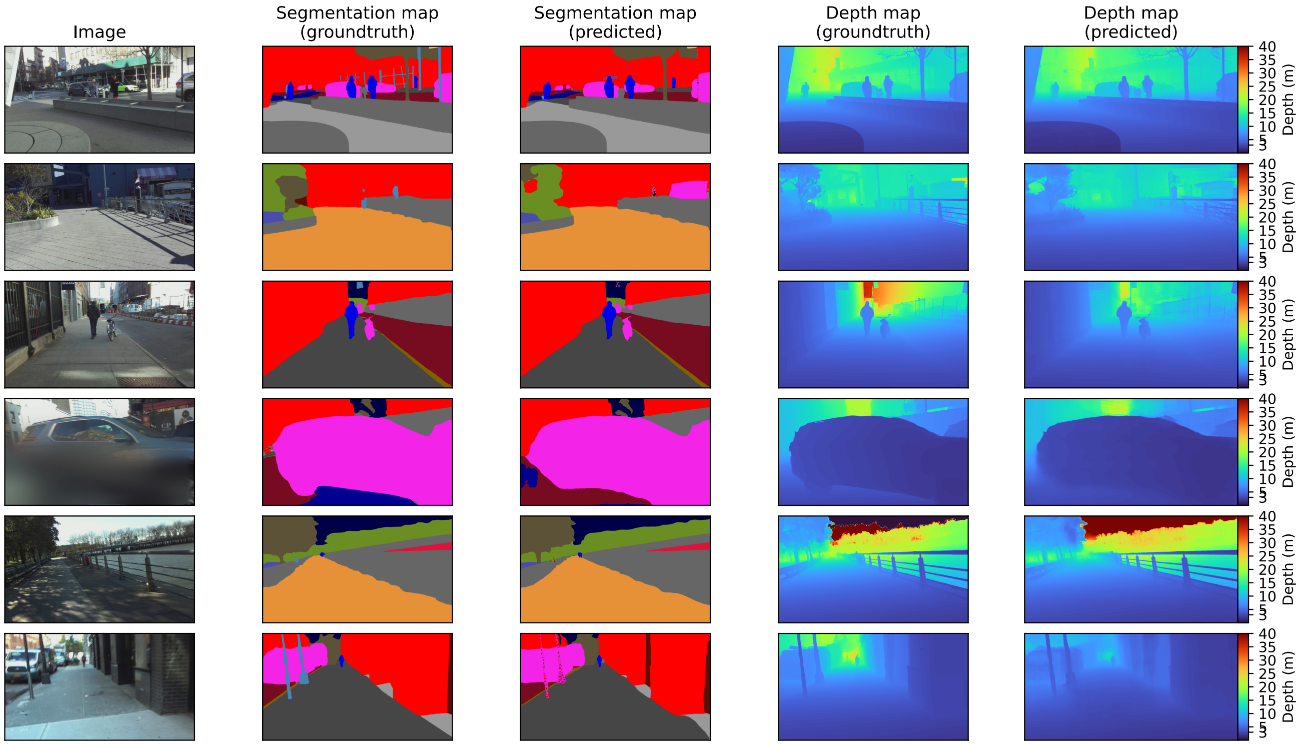

Figure 15. **Qualitative examples on SANPO.** Showing left to right: image, groundtruth & predicted segmentation maps, and groundtruth & predicted metric depth maps.

13. pedestrian → obstacles
14. rider → obstacles
15. animal → obstacles
16. stairs → safe to walk
17. water body → not safe to walk
18. other walkable surface → safe to walk
19. inaccessible surface → not safe to walk
20. railway track → not safe to walk
21. obstacle → obstacles
22. vehicle → obstacles
23. traffic sign → obstacles
24. traffic light → obstacles
25. pole → obstacles
26. bus stop → obstacles
27. bike rack → obstacles
28. sky → not safe to walk
29. tree → obstacles
30. vegetation → obstacles
31. terrain → safe to walk